

# The prediction of topologically partitioned intra-atomic and inter-atomic energies by the machine learning method kriging

Peter Maxwell<sup>1,2</sup> · Nicodemo di Pasquale<sup>1,2</sup> · Salvatore Cardamone<sup>1,2</sup> · Paul L. A. Popelier<sup>1,2</sup>

Received: 24 May 2016 / Accepted: 11 July 2016 / Published online: 27 July 2016  
© The Author(s) 2016. This article is published with open access at Springerlink.com

**Abstract** The construction of a novel protein force field called FFLUX, which uses topological atoms, is founded on high-rank and fully polarizable multipolar electrostatics. The machine learning method kriging successfully predicts multipole moments of a given atom with as only input the nuclear coordinates of the atoms surrounding this given atom. Thus, trained kriging models accurately capture the polarizable multipolar electrostatics of amino acids. Here we show that successful kriging models can also be constructed for non-electrostatic energy contributions. As a result, the full potential energy surface of the (molecular) system trained for can be predicted by the corresponding set of atomic kriging models. In particular, we report on the performance of kriging models for each atom's (*A*) (1) total atomic energy ( $E_{\text{IQA}}^A$ ), (2) intra-atomic energy ( $E_{\text{intra}}^A$ ) (both kinetic and potential energy), (3) exchange energy ( $V_{\text{XC}}^{AA'}$ ) and (4) electrostatic energy ( $V_{\text{el}}^{AA'}$ ) of atom *A* with the rest of the system (*A'*), and (5) interatomic energy ( $V_{\text{inter}}^{AA'}$ ). The total molecular energy can be reconstructed from the kriging predictions of these atomic energies. For the three case studies investigated (i.e. methanol, *N*-methylacetamide and peptide-capped glycine), the molecular energies were produced with mean absolute errors under 0.4, 0.8 and 1.1 kJ mol<sup>-1</sup>, respectively.

**Keywords** Quantum chemical topology (QCT) · Interacting quantum atoms (IQA) · Quantum theory of atoms in molecules (QTAIM) · Kriging · Machine learning · Amino acids · Force field

## 1 Introduction

There is a consensus that traditional force fields, which are instrumental in the vast majority of modern molecular simulations, need further improvement. Their limiting accuracy is regularly pointed out in the literature, as a cause for discrepancies between a given force field's predictions and experimental results. Indeed, if the sampling during the simulation is adequate, then only the potential can be blamed if predictions fail to be trustworthy. In order to improve their energy prediction, popular force fields such as AMBER and CHARMM have been modified on several occasions. Fairly recent modifications were extensively tested [1] in 2011 with the most powerful dedicated molecular simulation hardware in the world. The four force fields tested in this protein folding work were Amber ff03, Amber ff99SB\*-ILDN, CHARMM27 and CHARM22\*. It was found that the folding mechanism and the properties of the unfolded state depended substantially on the choice of force field. Another extensive and more recent study [2] concluded, from a millisecond of simulations on intrinsically disordered proteins, that eight well-known force fields generate unexpectedly huge differences in chain dimension, hydrogen bonding and secondary structure content. In fact, discrepancies are so serious that changing the force field has a stronger effect on secondary structure content than changing the entire peptide sequence. Such comparisons are quite rare but precious because they clearly demonstrate, while eliminating any sampling issues or hardware

**Electronic supplementary material** The online version of this article (doi:10.1007/s00214-016-1951-4) contains supplementary material, which is available to authorized users.

✉ Paul L. A. Popelier  
paul.popelier@manchester.ac.uk

<sup>1</sup> Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester M1 7DN, UK

<sup>2</sup> School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, UK

limitations, that much more work needs to be done. The question is which type of work.

Over the last decade our strategy has been to re-examine and challenge the core architecture of classical force fields. The type of work that accompanies such a bold strategy can be characterized as arduous, and above all, systematic. At the outset of this long term project, the ubiquitous atomic point charges (one on each nucleus) were replaced by nucleus-centred atomic multipole moments. This step is shared by other next-generation force fields such as AMOEBA [3], XED [4], SIBFA [5] and ACKS2 [6] and is driven by a clearly justified desire towards increasingly accurate electrostatics [7, 8].

The current approach embraces so-called topological atoms as the “entity of information” from which any system (molecular or ionic) is built. Much work has been carried out [9–17] in order to obtain a deep understanding of the convergence behaviour and accuracy of the electrostatic interaction between topological atoms, as well as the electrostatic potential they generate. Topological atoms are defined by the quantum theory of atoms in molecules (QTAIM) [18–21] as finite-volume fragments in real 3D space. As quantum atoms [20, 22], topological atoms are deeply rooted in quantum mechanics [23]. These atoms result from a parameter-free partitioning of the electron density, introducing sharp boundaries whose shape responds to any variation in nuclear geometry. The finite size of topological atoms prevents penetration effects and the associated correction in the form of damping functions. Topological multipolar electrostatics proved to be successful in the description of electrostatic interaction in proteins [24].

The next step in the construction of a topological force field is the inclusion of electrostatic polarization. In principle, the multipole moments of any given atom are influenced by all the atoms surrounding it, but this influence typically decreases the further away the surrounding atoms are. In order to be able to handle the full complexity of this influence, we invoked machine learning early on. Initially we used neural networks [25] and applied it to water clusters [26]. In 2009 it turned out [27] that a completely different machine learning method called kriging performed more accurately than neural networks. Although kriging was computationally more expensive, it coped better with the larger number of molecules surrounding the atom of interest. Kriging [28], which is also known as Gaussian regression analysis [29], originates in geostatistics but has been used in very different application areas, including the prediction [30] of atomic properties when inside molecules.

The essence of our kriging approach is the establishment of a direct mapping between an atomic multipole moment (output) and the nuclear coordinates (input) of the surrounding atoms. This mapping is obtained after training to a

training set of (molecular or cluster) geometries, and is able to make a prediction to a previously unseen geometry of the surrounding molecules. For this purpose we take advantage of the renowned interpolative power of kriging. Moreover, when using kriging models in an extrapolative way, the model returns the average value of the atomic property of interest observed over the training set. Finally, it must be pointed out that a kriging model is not returning an atomic polarizability but the atomic multipole moment itself, after the polarization process is complete. This strategy has an important advantage when the kriging models are used during a molecular dynamics simulation: *the atomic moments do not need to be computed (typically iteratively) from the polarizabilities*. Instead, the multipole moments are predicted “on the fly”, directly from the nuclear coordinates of the surroundings at any given time step.

Most attention has been devoted to modelling the electrostatic interaction at long-range by means of kriged atomic multipole moments. As this procedure is understood and works well [31–37], the next step is how to combine this electrostatic energy with the non-electrostatic energy contributions. Preliminary and unpublished work expressed the latter in the traditional manner, i.e. with Hooke-like potentials reinforced with anharmonic extensions. The parameterization of these potentials, in the presence of kriged electrostatics, turned out to be inadequate. For this reason, a more satisfactory and elegant alternative strategy was carried out, which is to combine kriged electrostatics with kriged non-electrostatics. In this streamlined procedure the machine learning method is trained for energy quantities that are obtained from the same topological energy partitioning [38] that yields the atomic multipole moments. In 2014, the atomic kinetic energy was successfully kriged [39] as the first non-electrostatic energy contribution. That work presented proof-of-concept based on four molecules of increasing complexity (methanol, *N*-methylacetamide, glycine and triglycine). For all atoms tested, the mean atomic kinetic energy errors fell below 1.5 kJ mol<sup>−1</sup>, and far below this value in most cases.

In the current article, we go further and deliver proof-of-concept for the kriging of non-electrostatic atomic energy contributions. For that purpose we have adopted the interacting quantum atoms (IQA) scheme proposed by Blanco et al. [40]. This is a topological energy partitioning scheme, inspired by early work [11] on atom–atom partitioning of intramolecular and intermolecular Coulomb energy. In IQA, the kinetic energy is subsumed in the intra-atomic energy (or sometimes called “self energy”), which also contains the potential energy of the electrons interacting with themselves and with the nucleus, both within a given atom. This intra-atomic energy plays a pivotal role in stereo-electronic effects, including intermolecular Pauli-like repulsion. Furthermore, IQA can calculate the electrostatic

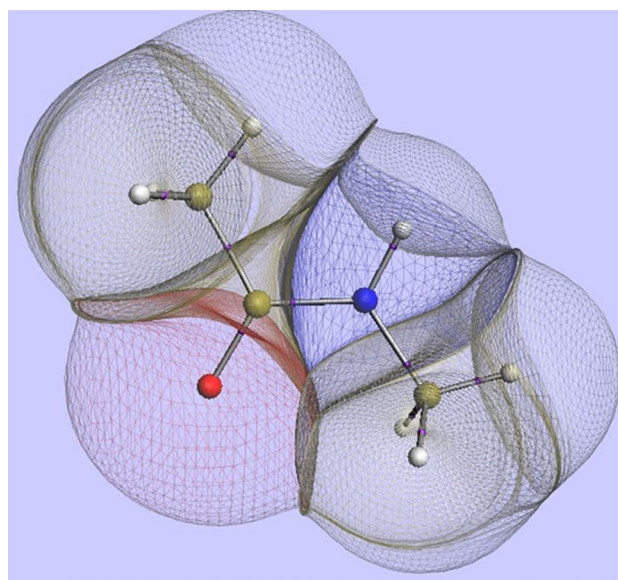
interaction between two atoms that are so close that their multipolar expansion diverges. The IQA method achieves this goal by using a variant of the six-dimensional volume integration (over two atoms) proposed in Ref. [11], which avoids the multipolar expansion altogether. Similarly, IQA does not multipole-expand the (inter-atomic) exchange energy, although this can be done [15]. However, this route is not followed by our topological force field. This energy contribution expresses covalent bonding energy. Within the Hartree–Fock *ansatz*, these three energy contributions (self, Coulomb and exchange) complete an IQA partitioning. However, post-Hartree–Fock methods introduce a fourth (non-vanishing) contribution that is associated with electron correlation. In the current work, we use a version of IQA [41] that is compatible with DFT with an eye on including electron correlation effects. We invoke the use of DFT level with the largest of systems studied here, capped glycine.

Three molecules were chosen here for this proof-of-concept investigation: methanol, *N*-methylacetamide (NMA) and glycine, which is capped by a peptide bond both at its C and N terminus. These systems were chosen to represent a progressive sequence of molecular complexity, while being relevant to biomolecular modelling: methanol features as the side residue in the amino acid serine, NMA is the smallest system modelling a peptide bond, while capped glycine represents an amino acid in an oligopeptide. This work is the first report of combining machine learning with a *full* topological energy partitioning.

## 2 Methodological background

### 2.1 The interacting quantum atoms (IQA) approach

Figure 1 shows examples of topological atoms appearing in *N*-methylacetamide, which were generated by in-house software [42, 43]. QTAIM essentially defines a topological atom as a three-dimensional subspace determined by the bundle of gradient paths (of a system's electron density) that are attracted to the atom's nucleus. This partitioning idea also features in other topological approaches [44], such as that in connection with the electron localization function (ELF). The topological energy partitioning method IQA is a third approach that uses the central idea of the so-called gradient vector field to extract chemical information from a system. Quantum chemical topology (QCT) [45] is a collective name to gather all topological approaches (10 so far, see Box 8.1 in Ref. [20]) that share the abovementioned central idea. The acronym QCT resurfaces in QCTFF, the force field under construction here [22], which uses topological atoms. The new name for QCTFF is FFLUX, for which a very recent and accessible perspective [46] can be



**Fig. 1** Topological atoms in a conformation of *N*-methylacetamide (NMA)

consulted. We also note here that it has been shown before [47] that atom types that can be computed using the atomic properties of topological atoms in amino acids.

It is clear from Fig. 1 that QCT partitions a molecule into well-defined non-overlapping atoms [48]. Moreover, these topological atoms do not show any gaps between them; in other words, they partition space exhaustively. It is important to pause and briefly discuss the full consequence of this property. Exhaustive partitioning infers that *each point in space* belongs to a topological atom: all space is accounted for. In principle, this property must have repercussions [49] for docking studies, as will become clear when QCT starts being used at this larger molecular scale. Classical drug design (e.g. [50]) thinks of both ligand and the protein's active site as bounded by artificial surfaces (e.g. solvent accessible surface) based on standard van der Waals radii and an image of overlapping hard spheres. This view necessarily introduces “gaps of open space”, which belong neither to the ligand nor to the protein. However, quantum mechanically we know that electron density resides in those gaps, no matter how small or thin they are. Electron density generates an electrostatic potential, and hence, also generates electrostatic interaction energy contributions. If a gap is not accounted for, then energy will be missing, which interferes with the energy balance during the docking process. However, if there is no gap, as in QCT, then all energy is properly accounted for.

In brief, IQA quantitatively describes the total energy of an atom, even if the system is not at a stationary point in the potential energy surface. In other words, unlike in orthodox QTAIM, there is no need to invoke the atomic virial theorem

[20, 51], the application of which is restricted to stationary points (such as equilibrium geometries). This total energy is comprised of the energy associated with the atom itself (intra-atomic), and with energy resulting from the interaction between the atoms (interatomic). We will explain each type of energy in turn, beginning with the decomposition of the molecular energy,  $E_{\text{IQA}}^{\text{molec}}$ , into the atomic energies, one for each atom  $A$ , denoted  $E_{\text{IQA}}^A$ , followed by its breakdown into intra-atomic (or ‘self’) and interatomic interaction energies,

$$E_{\text{IQA}}^{\text{molec}} = \sum_A E_{\text{IQA}}^A = \sum_A E_{\text{intra}}^A + \frac{1}{2} \sum_A \sum_{B \neq A} V_{\text{inter}}^{AB} \quad (1)$$

$$= \sum_A \left[ E_{\text{intra}}^A + \frac{1}{2} \sum_{B \neq A} V_{\text{inter}}^{AB} \right]$$

where  $E_{\text{intra}}^A$  and  $V_{\text{inter}}^{AB}$  are the *intra*-atomic (of atom  $A$ ) and *inter*-atomic (between atoms  $A$  and  $B$ ) energies, respectively. The intra-atomic energy can be further partitioned,

$$E_{\text{intra}}^A = T^A + V_{\text{ee}}^{AA} + V_{\text{en}}^{AA} \quad (2)$$

where  $T^A$  is the kinetic energy of the electrons associated with atom  $A$ ,  $V_{\text{ee}}^{AA}$  is the (repulsive) electron–electron potential energy, and  $V_{\text{en}}^{AA}$  is the (attractive) electron–nuclear potential energy. Together, these three energies comprise the intra-atomic energy possessed by a single atom.

The interatomic energy attributed to a pair of atoms can also be further partitioned,

$$V_{\text{inter}}^{AB} = \left( V_{\text{nn}}^{AB} + V_{\text{en}}^{AB} + V_{\text{ne}}^{AB} \right) + V_{\text{ee}}^{AB} \quad (3)$$

where  $V_{\text{en}}^{AB}$ ,  $V_{\text{ne}}^{AB}$  and  $V_{\text{ee}}^{AB}$  were described above but now with the ordering of the subscript components being allied to the ordering of the atoms in the superscript. For example, subscript ‘en’ and superscript ‘AB’ refers to the electrons of atom  $A$  and the nucleus of atom  $B$ . In addition to the electronic energy components,  $V_{\text{nn}}^{AB}$  is the repulsive nucleus–nucleus potential energy.

The electron–electron energy  $V_{\text{ee}}^{AB}$  can be even further partitioned to give the components in Eq. (4),

$$V_{\text{ee}}^{AB} = V_{\text{Coul}}^{AB} + V_X^{AB} + V_{\text{corr}}^{AB} \quad (4)$$

where  $V_{\text{Coul}}^{AB}$  represents the Coulombic interaction between the electrons in atoms  $A$  and  $B$ .  $V_X^{AB}$  represents the inter-electron exchange potential energy and  $V_{\text{corr}}^{AB}$  the inter-electron correlation potential energy. Combining the bracketed terms in Eq. (3) with the Coulomb energy only, leads to the total electrostatic energy between two atoms, or  $V_{\text{elec}}^{AB}$ , which is often written as  $V_{\text{cl}}^{AB}$  because of the ‘classical’ nature of the electrostatic potential energy (devoid from any purely quantum mechanical exchange energy),

$$V_{\text{cl}}^{AB} = \left( V_{\text{nn}}^{AB} + V_{\text{en}}^{AB} + V_{\text{ne}}^{AB} \right) + V_{\text{Coul}}^{AB} \quad (5)$$

We have extensively studied this energy  $V_{\text{cl}}^{AB}$  in terms of its multipolar convergence behaviour. The quantity  $V_{\text{cl}}^{AB}$  incorporates the widely reported electrostatic multipole moments’ contribution of the long-ranged electrostatic energy, in addition to the short-range electrostatic contribution obtained from IQA. Equation (3) can be rewritten as Eq. (6). This is done by first substituting the bracketed expression by  $V_{\text{cl}}^{AB} - V_{\text{Coul}}^{AB}$ , as obtained from Eq. (5), and then by substituting  $V_{\text{ee}}^{AB}$  using Eq. (4), such that after cancellation of  $V_{\text{Coul}}^{AB}$  we obtain,

$$V_{\text{inter}}^{AB} = V_{\text{cl}}^{AB} + V_X^{AB} + V_{\text{corr}}^{AB} = V_{\text{cl}}^{AB} + V_{\text{XC}}^{AB} \quad (6)$$

The new expression separates the interatomic energy into the interplay of ionic-like ( $V_{\text{cl}}^{AB}$ ), covalent ( $V_X^{AB}$ ) and correlation ( $V_{\text{corr}}^{AB}$ ) energies. Note that it is often convenient to combine exchange and correlation in one term. These three energies along with the intra-atomic energy compose the four primary energies that FFLUX is built upon.

Until recently [41] the inclusion of any computationally affordable correlation energy has been lacking because IQA is incompatible with both perturbation theory and density functional theory (DFT) methods. Indeed, the methods that are compatible with the original IQA (i.e. full configuration interaction (FCI), complete active space (CAS), configuration interaction with single and double excitations (CISD) or coupled cluster with single and double excitations (CCSD) levels of theory) demand much greater computational expense. Neither perturbation theory nor standard DFT methods provide a well-defined second-order reduced density matrix, and hence IQA cannot be straightforwardly applied to them. Together with Dr TA Keith, the main author of the QCT computer program AIMALL [52], a DFT-based IQA method that incorporates at least some correlation was validated [41] by our group. The solution involved incorporating the explicit B3LYP atomic exchange functional in order to correctly calculate an atom’s *total* atomic exchange, thereby recovering the ab initio energy of the whole molecule. However, the fact that the functional cannot be used to calculate interatomic exchange (see Ref. [41] for details) led us to calculate the interatomic component using the Hartree–Fock-like expression but then with Kohn–Sham orbitals inserted. The remaining intra-atomic exchange–correlation is then calculated as the difference between the atomic exchange–correlation directly obtained from the B3LYP functional and the Hartree–Fock-like interatomic exchange.

In this investigation, we will again make use AIMALL. This program is able to return a useful quantity, denoted  $V_{\text{inter}}^{AA'}$ , which is defined as follows:

$$V_{\text{inter}}^{AA'} = \sum_{B \neq A} V_{\text{inter}}^{AB} \quad (7)$$



where  $A'$  represents every atom other than atom  $A$ . Note that Eq. (7) defines the *atom-centred* interatomic energy contribution  $V_{\text{inter}}^{AA'}$ . In other words, it summarizes how an atom interacts in full with all other atoms. Note that the quantity  $V_{\text{inter}}^{AA'}$  is obtained computationally cheaper than by summing over the individual pair-wise atomic contributions  $V_{\text{inter}}^{AB}$ . However, the computational advantage of  $V_{\text{inter}}^{AA'}$  is offset by a reduction in the chemical insight that we obtain from being able to inspect each atom pair individually. This loss occurs over and above that caused by lumping together the electrostatic, exchange or correlation energy contributions (see Eq. 6). Instead, the interaction energy is defined in terms of a given atom  $A$  experiencing the entire surrounding molecular environment. Equally,  $V_{\text{inter}}^{AA'}$  can be decomposed into  $V_{\text{cl}}^{AA'}$  and  $V_X^{AA'}$  components, much like the pair-wise  $AB$  energies. For our current purpose, we are satisfied with this formulation in spite of the reduced chemical insight it gives because our prime motivation is to predict atomic energies, and not to predict chemical insight.

Returning to the IQA formalism, we obtain Eq. (8), which expresses three ways (“approaches”) to break up the molecular energy into atomic contributions. Approach A was already present in Eq. (1) while substituting  $V_{\text{inter}}^{AA'}$  of Eq. (7) into Eq. (1) leads to Approach B. Finally, Approach C follows from Eq. (6) and now applying the idea behind Eq. (7) to  $V_{\text{cl}}^{AB}$  and  $V_{\text{XC}}^{AB}$  yielding

$$\begin{aligned}
 E_{\text{IQA}}^{\text{molec}} &= \underbrace{\sum_A E_{\text{IQA}}^A}_{\text{Approach A}} = \underbrace{\sum_A E_{\text{intra}}^A + \frac{1}{2} \sum_A V_{\text{inter}}^{AA'}}_{\text{Approach B}} \\
 &= \underbrace{\sum_A E_{\text{intra}}^A + \frac{1}{2} \sum_A V_{\text{XC}}^{AA'} + \frac{1}{2} \sum_A V_{\text{cl}}^{AA'}}_{\text{Approach C}} \quad (8)
 \end{aligned}$$

which is the key equation for the analysis in this paper. Note that  $V_{\text{inter}}^{AA'}$  is always halved when used in Eq. (8), attributing half of the energy to a single atom  $A$ , in order to prevent double-counting of the interatomic energy in the molecule.

We aim for a greater understanding of both the quantitative nature of these five types of energy ( $E_{\text{IQA}}^A$ ,  $E_{\text{intra}}^A$ ,  $V_{\text{inter}}^{AA'}$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ ) and their suitability in FFLUX after being kriged. As suggested in Eq. (8), the molecular energy can be recovered through three different approaches, each incorporating the use of different IQA energies. Approach A uses only the total atomic energy of an atom, denoted  $E_{\text{IQA}}^A$ . The atomic energy  $E_{\text{IQA}}^A$  is a sum of the intra- and inter-atomic energy and hence expresses their resulting “trade-off” by the single quantity that it is. This final energy,  $E_{\text{IQA}}^A$ , suffices by itself for FFLUX being able to predict the structure and dynamics of a system, because the latter only depend on the total atomic energy, not its

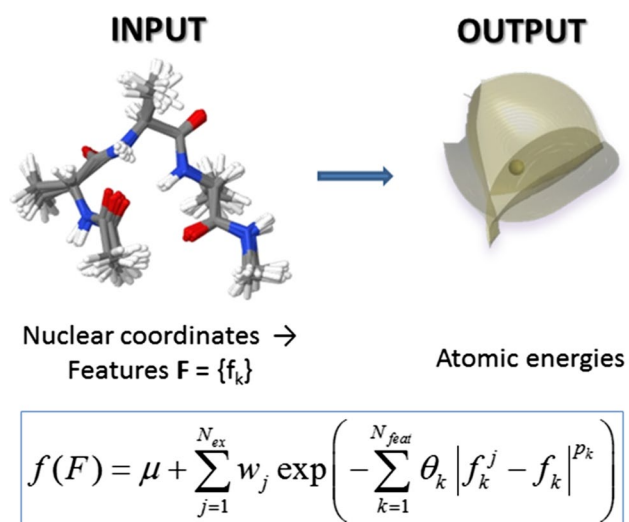
breakdown. Approach B exposes the separation of the intra-atomic and interatomic energies for insight into how an atom itself experiences the environment it is in. Finally, Approach C takes the separation one step further, using the individual exchange and electrostatic energies in the interatomic description of an atom.

In order to clarify the strategy for the *complete* treatment of energy contributions in FFLUX a comment about  $V_{\text{corr}}^{AB}$  in Eq. (4) is necessary. This energy contribution covers dynamic correlation and hence dispersion. Our preferred route is to treat  $V_{\text{corr}}^{AB}$  in exactly the same way as  $V_{\text{cl}}^{AB}$  and  $V_X^{AB}$ . This approach, for which proof-of-concept has been reached in our lab, will guarantee a seamless integration of dispersion in the FFLUX *ansatz*. This strategy will thereby avoid the typical problems (e.g. the need for damping functions) that alternative dispersion methods introduce.

For a more exhaustive description of the IQA partitioning scheme including additional formulae, its capabilities and previous applications, we refer to the original literature [40, 53–58].

## 2.2 Kriging (Gaussian regression analysis)

As a machine learning method, kriging has its roots in geostatistics where it has been used to *predict* the location of precious material after being *taught* these locations [28]. Within FFLUX, kriging is used to map geometrical change within a molecule, obtained from nuclear coordinates, to a corresponding atomic property, which can be an IQA energy or atomic multipole moment. The atomic property is the machine learning output and the coordinates are the input. Although the full details are given elsewhere [33, 34] we explain here how these coordinates are constructed. It is advantageous that the coordinates are internal in nature (so only  $3N - 6$  for a nonlinear  $N$ -atom system). On each nucleus we install a so-called atomic local frame (ALF), which enables the definition of a polar angle and an azimuthal angle to describe the position of each nucleus in the system, except for the three nuclei required in defining the ALF. The distance between the ALF's origin and the nucleus completes the triplet of (spherical polar) coordinates for a given nucleus in the system (other than that on which the ALF is installed). Machine learning language calls these coordinates *features*, as they are the input variables to kriging in this case. Finally we note that the first three features of the vector of  $3N - 6$  features (necessary to describe unambiguously a molecular geometry) consists of (1) the distance between the origin and the first nucleus, which fixes the ALF's  $x$ -axis, (2) the distance between the origin and the second nucleus (fixing the ALF's  $xy$ -plane), and (3) the angle suspended by the first nucleus, the origin and the second nucleus.



**Fig. 2** Summary of kriging method at the heart of FFLUX. Atomic energies (intra-atomic, inter-atomic, or total sum) of a given topological atom (*right panel*, output) are mapped onto the features  $\{f_k\}$ , which describe the nuclear geometry of the environment surrounding this given atom. The kriging parameter  $\theta_k$  and  $p_k$  are optimized (see main text)

Before introducing the mathematics involved, it is useful to understand the kriging procedure qualitatively. Before a kriging model can predict a quantity, it must first be trained for using a number of molecular geometries with corresponding atomic properties. These data form the training set while the test set will consist of data that do *not* belong to the training set. The external character of the test set makes the assessment of the kriging predictions meaningful, because predictions of the training set data would be exact anyway (with the type of kriging used here). The molecular geometries used to build each set are obtained from sampling a molecular energy well that surrounds an energy minimum. The details of this sampling are given below, in Sect. 2.3.

Figure 2 summarizes the kriging approach, showing the key formula that maps the input (left panel, features) to the output (right panel, atomic energies). The feature vector  $\mathbf{F}$  collects all  $N_{\text{feat}}$  features  $f_k$ , which have been detailed above. Note that  $N_{\text{feat}}$  is the dimensionality of the *feature space* in which all  $N_{\text{ex}}$  molecular geometries are expressed. In the formula of Fig. 2,  $N_{\text{ex}}$  is the number of training examples,  $\mu$  represents the mean of the observed value (also known as a constant “background” value) while  $w_i$  is the kriging “weight” (obtained from the so-called correlation matrix [33]). Note that the name weight should not conjure up an image of arbitrary adjustments because each weight is computed exactly as explained in Ref. [33].

Again, instead of giving the full mathematical details [33, 59] of the kriging procedure, we here explain the crux in words. Imagine a coin being tossed a number of times

and the outcomes recorded. If the coin is fair, then the parameter, which governs the outcomes and which is called  $t_h$ , is exactly 0.5 for tossing a head. This parameter is analogous to parameters  $\theta_k$  and  $p_k$  in the formula of Fig. 2. Now suppose that we observe a statistical bias towards the outcome of “heads up”. We can then ask to what the extent the coin is biased. In other words, given the observations made  $t_h$  differs from 0.5. In fact, one can find the value of the parameter  $t_h$  such that the likelihood is maximal of again observing the outcomes that were observed. For example, one could find that  $t_h = 2/3$  is this value. Similarly, when this idea is applied to the kriging problem at hand, the so-called likelihood function is maximized against the observed data, which are the atomic energies. This procedure then returns the optimal values of the parameters  $\theta_k$  and  $p_k$ . The technical details of this optimization are complex and extensively researched in our laboratory [59]. The next paragraph provides a very brief summary. With regard to using the key kriging formula in Fig. 2 (after training), previously unseen features  $\mathbf{F} = \{f_k\}$  are inserted, returning the atomic property  $f(\mathbf{F})$ . It is clear that the argument of the exponential is a distance function, which is not necessarily Euclidean (i.e.  $p_k \neq 2$ ).

In terms of the optimization procedure, first the concentrated log-likelihood is calculated analytically. The function is then maximized by a different machine learning method because this cannot be achieved analytically. We have successfully used particle swarm optimization (PSO), the mathematical details of which can be found in Ref. [60]. The optimization of the parameters  $\theta_k$  and  $p_k$  via PSO is the most computationally expensive step in the overall kriging process. However, optimizing these parameters allows the user to obtain the highest possible concentrated log-likelihood function, ensuring that the best possible model is obtained. The time for the PSO optimization is proportional to the number of geometries in the training set and the number of atoms in the molecule. The result is an analytical formula (see Fig. 2) linking the geometrical features of a molecule and the atomic property of choice. For a more comprehensive description of the kriging protocol, the reader is invited to refer to our previous publications [33, 35–37, 59, 61].

### 2.3 Sampling of distorted geometries

The selection of training examples with which to construct a kriging model is of great importance. The geometries of the training set should be representative of the physically realistic regions of conformational space. This representation ensures that predictions corresponding to relevant molecular geometries are always made in areas of conformational space that have been trained for in the kriging model. Conventional methods will use some form of

molecular mechanics, parameterized by a classical force field, to generate the training geometries. However, it has been shown that molecular mechanics does not necessarily sample the relevant areas of conformational space over the course of a typical trajectory [62–64].

Our approach attempts to locally approximate the *ab initio* molecular potential energy surface about a “seeding” conformation, which in this work is the global energetic minimum of the molecule. Whilst more than one seeding conformation can be used, granting a greater exploration of the potential energy surface, we found that for the molecules considered in this work, the Boltzmann weight of the global minimum exceeded 0.75 in all cases. By evaluating the first- and second-order spatial derivatives of the potential energy (Jacobian and Hessian, respectively) at the seeding geometry, one can construct an approximate local potential energy surface through a Taylor expansion. The dynamics on this local approximation to the potential energy surface are then governed by a set of harmonic equations of motion, referred to as the molecular normal modes [65].

Here we outline the major features of our methodology, whilst a more thorough description of is given elsewhere [66, 67]. For an  $N$ -atom molecular system, we can define a  $3N \times 3N$  transformation matrix,  $\mathcal{D}$ , that converts a mass-weighted Cartesian state vector,  $\mathbf{q}$ , to an internal coordinate state vector,  $\mathbf{s}$ . Given  $\mathcal{D}$ , we can transform the mass-weighted Cartesian Hessian,  $\mathbf{H}_q$ , to an internal coordinate basis through

$$\mathcal{D}^T \mathbf{H}_q \mathcal{D} = \mathbf{H}_s \quad (9)$$

The frequencies of the molecular normal modes are then given by diagonalising  $\mathbf{H}_s$

$$\mathcal{E}^{-1} \mathbf{H}_s \mathcal{E} = \mathbf{I} \lambda \quad (10)$$

where  $\mathcal{E}$  correspond to the eigenvectors of  $\mathbf{H}_s$ ,  $\mathbf{I}$  is the identity matrix, and the eigenvalues of  $\mathbf{H}_s$  are given by the  $3N$  diagonal elements  $(\mathbf{I} \lambda)_{ii} = \lambda_i$ . The  $i$ th normal mode frequency,  $\nu_i$ , is related to the  $i$ th eigenvalue through the expression

$$\nu_i = \sqrt{\frac{\lambda_i}{4\pi^2 c^2}} \quad (11)$$

where  $c$  is a conversion factor incorporating the speed of light and a conversion from atomic units to reciprocal centimetres. Six of these normal mode frequencies are equal to zero in an internal coordinate basis, corresponding to the global translational and global rotational degrees of freedom of the molecular system.

The amplitude of vibration of the  $i$ th normal mode,  $A_i$ , is given by the standard expression for a simple harmonic oscillator,

$$A_i = \sqrt{\frac{2E_i}{k_i}} \quad (12)$$

where  $k_i$  is the force constant of the  $i$ th normal mode (calculable from  $\nu_i$ ), and  $E_i$  is the energy available to it. Each normal mode is allocated an amount of energy given by a standard equipartition,  $E_i = k_B T/2$ , where  $k_B$  is the Boltzmann constant and  $T$  is the temperature at which the simulation is performed. To allow for a little more flexibility, each of the  $E_i$  is subjected to a stochastic Gaussian fluctuation. Given the amplitude and frequency of the oscillator, each normal mode can evolve in discrete time.

The final matter requiring discussion is how time is discretized for our equations of motion. Given the frequency of the oscillator, we can compute its time period,  $T_i = 1/\nu_i$ . We choose a discrete timestep,  $\Delta t_i$ , such that for each time period,  $T_i = \Delta t_i n_{\text{cycle}}$ , where  $n_{\text{cycle}}$  is a user-defined parameter. After every  $n_{\text{cycle}}$  timesteps, we also perturb the energy available to each normal mode by a new Gaussian-distributed number. To reduce the correlation between samples, a final user-defined parameter,  $n_{\text{out}}$  is also defined. We define  $n_{\text{out}}$  to correspond to the number of discrete timesteps that we allow to pass before outputting a sample to the training set. For the work conducted here, we set  $n_{\text{cycle}} = 10$  and  $n_{\text{out}} = 100$ .

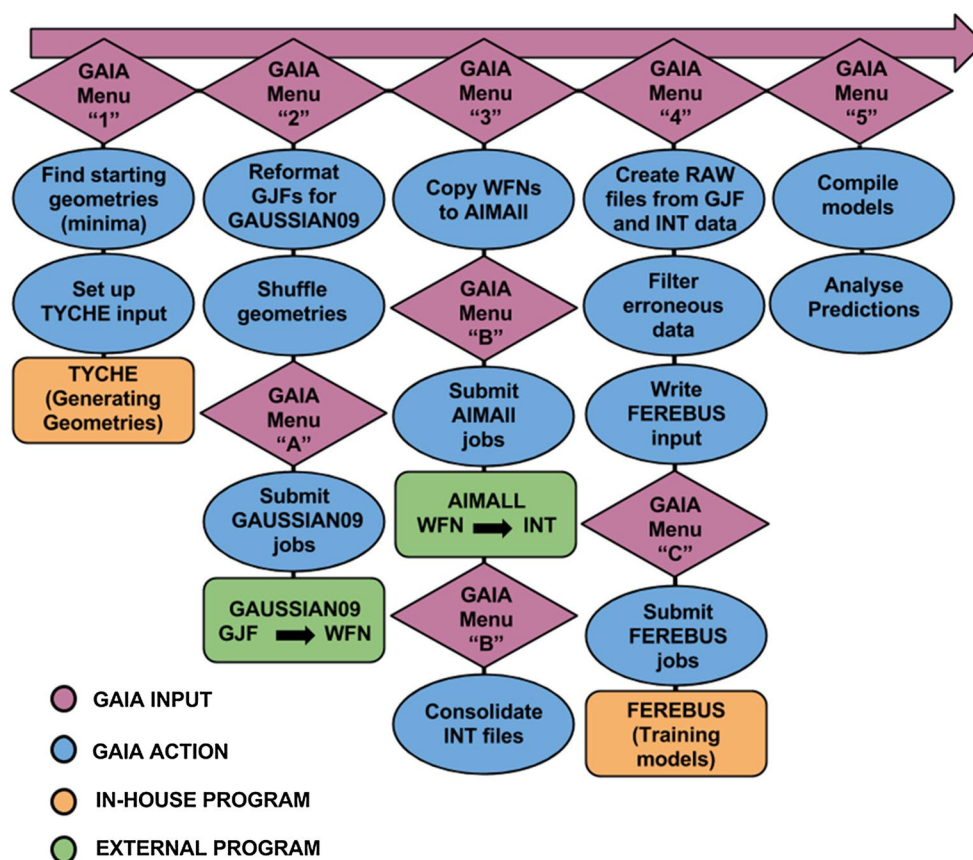
### 3 Computational methods

#### 3.1 The GAIA protocol

Three molecules (methanol, NMA and peptide-capped glycine) have been selected as case examples. Initially, the methanol and NMA molecules were generated in Gauss-View and optimized to a minimum energy geometry using the Gaussian 09 program, at HF/6-31+G(*d,p*) level of theory. Single-point energy calculations were performed on the resulting structures, outputting the respective molecular wavefunction and Hessian of the potential energy, calculated at the same level. For the capped glycine molecule, we selected the global minimum conformation from the nine known energetic minima described in a previous publication [68]. For glycine, the calculations were performed at B3LYP/apc-1 [69] level of theory, in-keeping with the level of theory used in previous research [68]. Working at B3LYP level complements a recent publication [41] that validates the extension of the IQA approach to the B3LYP density functional. Prior to such work, the typical IQA partitioning restrictions demanded a well-defined second-order density matrix, thus ruling out correlation-inclusive and approximate Hamiltonian theories, including the density functional theory (DFT) functionals [40, 70–72].

After obtaining the molecular wavefunctions, the process of sampling, performing the energy partitioning and building the kriging models was achieved using the in-house pipeline software, known as the GAIA protocol. The

**Fig. 3** GAIA protocol used to develop kriging models for FFLUX



GAIA protocol is outlined in Fig. 3, which displays the sequence of steps, flowing left to right, starting with development of sample geometries and terminating with analysing the models.

GAIA automates the passing of information between two in-house Fortran 90 programs (TYCHE and FEREBUS—orange boxes in Fig. 3) and two commercially available programs (GAUSSIAN09 [73] and AIMAll [52]—green boxes). The output data from one step subsequently forms the input for the following step, until a seeding geometry (or set of seeding geometries) has been converted into a fully trained kriging model. Each program's role within GAIA can be summarized in a few lines:

1. TYCHE: distorts an input seed geometry, using the molecular normal modes, to create a broad range of sample geometries that collectively describe a local patch on the molecular potential energy surface (around the seed).
2. Gaussian 09 [73]: performs single-point energy calculations and outputs the wavefunction of each molecule.
3. AIMAll [52] (version 15.09.12): starts from the wavefunction of a molecule to obtain the IQA energy partition values:  $E_{\text{IQA}}^A$ ,  $E_{\text{intra}}^A$ ,  $V_{\text{inter}}^{AA'}$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ .

4. FEREBUS: uses a *training set* of molecular geometries to build kriging models of atomic energy (any of the five types above). FEREBUS then validates each model by predicting a *test set* and comparing the models' predicted value to the known true value.

The GAIA protocol outlined here is a slight deviation to that reported [22] before for the parameterization procedure of FFLUX. The deviation is a result of the current exclusion of atomic multipole moments but incorporation of the IQA atomic energy components instead. Thus, Fig. 3 represents the protocol tailored to this investigation only.

A set of 4000 initial samples were generated for each molecule by TYCHE from the distortion of a single energy minimum, at a user-defined temperature of 450 K. After single-point energy calculations and wavefunctions were obtained from Gaussian 09 for every sample, IQA energy contributions were obtained from AIMAll (with default quadrature and integration grid options). We set to the value of 3 the AIMAll parameter '*-encomp*' referring to the IQA energies to be computed. As soon as one atom attains a Lagrangian integration error,  $L(\Omega)$ , greater than the user-defined threshold of 0.001 Hartree, then this atom is removed from the training set, as well as all remaining atoms of the molecular geometry in which the offending



atom occurred. This process is known as *scrubbing* in GAIA. Scrubbing ensures that samples with “noisy” atomic energies (i.e. large  $L(\Omega)$  value) are excluded from the development of the model. From the samples remaining in the pool after scrubbing, 500 are set aside as the *test set* and the remaining number of samples (to the nearest hundred) are used for the *training set*. The resulting training sets were 3400, 3300 and 3000 for methanol, NMA and capped glycine, respectively. These training sets were then employed to generate kriging models for each molecule using the in-house program FEREBUS. The kriging parameter,  $p_k$ , was optimized in the development of all the models. The settings in FEREBUS were as follows: noisy kriging was not requested, tolerance set to  $10^{-9}$ , convergence to 200 and the swarm-specifier to “dynamic”. Finally, so-called *S*-curves are produced to illustrate the energies errors on each molecular model. The development and meaning of an *S*-curve is described in the next section.

### 3.2 Energy error analysis

As announced earlier, each molecule will be modelled using three approaches, resulting in a tiered level of chemical information being incorporated into the molecular models:

- Approach A: Modelling the molecule using only the total unpartitioned atomic energy,  $E_{\text{IQA}}^A$ .
- Approach B: Modelling the molecule using the intra-atomic ( $E_{\text{intra}}^A$ ) and interatomic ( $V_{\text{inter}}^{AA'}$ ) atomic energies.
- Approach C: Modelling the molecule using the intra-atomic energy ( $E_{\text{intra}}^A$ ) and the two key interatomic energies: exchange–correlation ( $V_{\text{XC}}^{AA'}$ ) and classical electrostatic ( $V_{\text{cl}}^{AA'}$ ).

Approach A provides the fastest (computationally) and simplest model of a molecule, at the atomic level. Approach B offers a chemically intuitive separation of the intra-atomic and *overall* interatomic energies and provides models for both. Approach C offers the highest level of chemical detail (in this investigation), separating the interatomic interaction energy into the covalent-like exchange and ionic-like electrostatic components, and again returns models for each.

Moving on to the analysis of the models, we should reintroduce how *S*-curves can be used to fully convey the quality of a kriging model. The *S*-curve is a cumulative distribution function (up to 100 %) of absolute energy errors for each test point within the test set. An *S*-curve plots the *absolute (energy) error over the whole molecule* (*x*-axis) versus the *test set data point* (i.e. molecular geometry) (*y*-axis) as represented as a percentage (100 %/500 data points = 0.2 % per data point). Thus, each test set

molecular geometry point corresponds to one point on the *S*-curve. In order to plot the total molecular energy error (*x*-axis), it must be calculated the generalized expression appearing in Eq. (13),

$$\Delta E_{\text{IQA}}^{\text{Molec}} = \left| \sum_A^{N_{\text{atoms}}} \sum_Y^{n \leq 3} [E_{Y,\text{Act}}^A - E_{Y,\text{Pred}}^A] \right| \quad (13)$$

where ‘Y’ is a general notation representing any of five possible IQA energy contributions ( $E_{\text{IQA}}^A$ ,  $E_{\text{intra}}^A$ ,  $V_{\text{inter}}^{AA'}$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ ), and  $n$  is the number of atomic energies being used to describe an atom (or the total atomic model). In this work  $n$  can be one, two or three only (hence the upper limit  $n \leq 3$  in Eq. 13). The value of  $n$  depends on the modelling approach. In particular, for approach A we have that  $n = 1$  ( $E_{\text{IQA}}^A$ ), for approach B  $n = 2$  ( $E_{\text{intra}}^A$  and  $V_{\text{inter}}^{AA'}$ ) and approach C leads to  $n = 3$  ( $E_{\text{intra}}^A$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ ). Before summing over the atoms, counted by index A, a sum over  $n$  atomic energies must take place in order to obtain the *atomic* model. When a model is tested, the predictions of the model can be averaged and compared to the average value of the true values. As a result, a model can be said to, on average, slightly over- or under-predict as determined by a positive or negative difference between the averaged predicted and true values. Therefore, summing over these energy models allows for a cancellation of such errors across the models in two possible ways: (1) across the atomic energies that together constitute a single atom, which is *atomic cancellation*, and (2) across the atomic models that together constitute the molecule, *molecular cancellation*. The value obtained for  $\Delta E_{\text{IQA}}^{\text{Molec}}$ , as plotted on the *S*-curve *x*-axis, represents the final error for the molecular energy.

The mean absolute error (MAE) for the molecular model is calculated according to Eq. (14). The MAE can be used as a simple measure of the model quality and can be calculated for a single energy model or for a collection of models (such as the resulting molecular model, see Eq. 13),

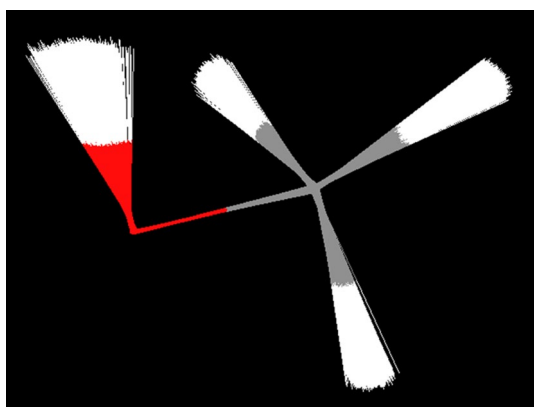
$$\Delta E_{\text{MAE}}^{\text{Molec}} = \frac{1}{N_{\text{test}}} \sum_{M=1}^{N_{\text{test}}} \Delta E_{\text{IQA},M}^{\text{Molec}} \quad (14)$$

where the sum runs over the  $N_{\text{test}} = 500$  molecular geometries of the test set (of methanol, NMA or glycine).

A final measure, the MAE percentage error, MAE%, can also be calculated by dividing the MAE by the size of the energy well range of the test set. This error is given in Eq. (15),

$$\text{MAE\%} = \frac{\Delta E_{\text{MAE}}^{\text{Molec}}}{E_{\text{MAX}}^{\text{TestSet}} - E_{\text{MIN}}^{\text{TestSet}}} \quad (15)$$

where ‘MAX’ refers to the highest molecular energy in the test set, and ‘MIN’ to the lowest. Note that the Electronic



**Fig. 4** Set of 4000 distorted methanol samples as generated from the in-house program TYCHE through sampling of the normal modes at a temperature of 450 K

Supplementary Material reports atomic MAE% values, which are defined in Eq. S1 there, by simply replacing  $\Delta E_{\text{MAE}}^{\text{Molec}}$  by  $\Delta E_{\text{MAE}}^{\text{Atom}}$ . Converting the error into a percentage allows a *transferable* measure, independent of the energy range of the sampling well, thereby making the MAEs from different molecules more comparable. The MAEs can also be used to compare the quality of individual atomic energy models (also in Eq. 15), which individually may experience a broad variety of energy ranges.

Finally, this work shows, for the first time, *S*-curves of the complete molecular energy ( $\Delta E_{\text{IQA}}^{\text{Molec}}$ ) rather than only the multipolar electrostatic energy or later the kinetic energy. Because multipole moments are not used in this work *all* atoms can interact with each other electrostatically (without concerns about possible divergence). In other words, the complete electrostatic interaction is subject to kriging here, for the first time, covering all 1,2; 1,3 and 1,4 interactions.

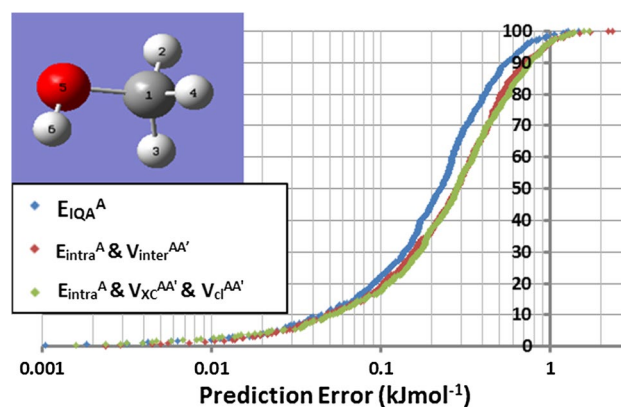
## 4 Results

### 4.1 Methanol

The 4000 samples (i.e. molecular geometries) generated by TYCHE (see Fig. 4) were subject to scrubbing in GAIA, followed by 500 samples then set aside as the test set. After rounding down to the nearest hundred, 3400 samples remained and formed the training set. These 3400 training set samples sampled an energy well with an energy range of  $\sim 115 \text{ kJ mol}^{-1}$ .

Figure 5 plots the *S*-curves for methanol for each of the three modelling approaches ( $A = E_{\text{IQA}}^A$ ,  $B = E_{\text{intra}}^A$  and  $V_{\text{inter}}^{AA'}$ ,  $C = E_{\text{intra}}^A$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ ).

It is clear that over 95 % of the test set geometries have  $\Delta E_{\text{MAE}}^{\text{Molec}}$  energy errors below  $1 \text{ kJ mol}^{-1}$ , across all



**Fig. 5** Methanol *S*-curves showing the absolute errors for each of the three modelling approaches, each tested on the same 500 test set samples

modelling approaches. Such a low error is a pleasing result and an encouraging start to this analysis. An analysis of each molecular model (i.e. approach) is given in Table 1. Interestingly, the simplest model, approach A, performs the best out of the three approaches with a MAE% error of 0.3 %. Approaches B and C perform very similarly with an MAE% error of 0.4 % each, respectively.

Notably, the maximum absolute error observed for approach B is  $2.3 \text{ kJ mol}^{-1}$ , larger than that for approach C, which returns  $1.7 \text{ kJ mol}^{-1}$ . One would expect that the most chemically insightful approach, which is C, is the one that accumulates the highest error. This presumption follows from the fact that approach C (which has 18 models, or 3 energies for each of the 6 atoms) has 6 additional models compared to approach B (with 12 models, or 2 energies for each of the 6 atoms), and the extra models accrue additional kriging errors with each model. This matter will be discussed in Sect. 5.2.

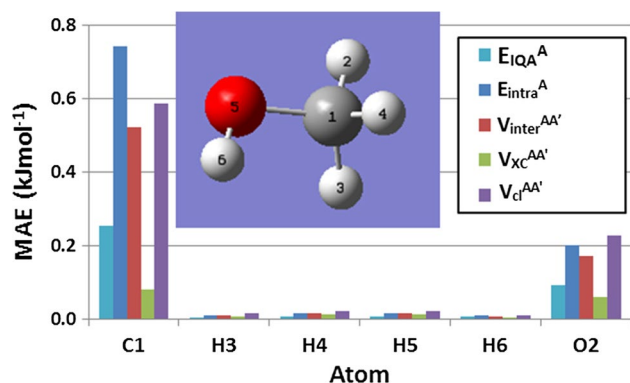
At the atomic level, Fig. 6 shows the MAEs for each atomic energy type ( $E_{\text{IQA}}^A$ ,  $E_{\text{intra}}^A$ ,  $V_{\text{inter}}^{AA'}$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ ) for each atom in methanol. At first glance, carbon influences the accuracy of the model. In general, C1 has the highest MAEs and is therefore the least accurately modelled atom overall. Following C1, O2 has the next highest errors, followed by the methyl hydrogens (H3/4/5), and finally the most accurately modelled atom, the alcoholic hydrogen H6. In assessing how the IQA atomic energy types compare, the story is also clear. Without exception, the following order appears, starting with the least accurate:  $E_{\text{intra}}^A$  &  $V_{\text{cl}}^{AA'} < V_{\text{inter}}^{AA'} < E_{\text{IQA}}^A < V_{\text{XC}}^{AA'}$ . The errors for all energies for any atom never exceed  $0.8 \text{ kJ mol}^{-1}$ . The interplay between these energies will be discussed in the Sect. 5.

Only looking at the MAE ignores the range of energy that a particular energy has been subjected to in the sampling stage. The MAE percentage error (MAE%) makes the

**Table 1** Quantitative analysis of the methanol models

| Measure                   | $E_{\text{IQA}}^A$ (approach A) | $E_{\text{intra}}^A$ & $V_{\text{inter}}^{A,A'}$ (approach B) | $E_{\text{intra}}^A$ & $V_{\text{XC}}^{A,A'}$ & $V_{\text{cl}}^{A,A'}$ (approach C) |
|---------------------------|---------------------------------|---|---|
| Maximum absolute error    | 1.5                             | 2.3   | 1.7   |
| Minimum absolute error    | 0.0                             | 0.0   | 0.0   |
| Absolute error range      | 1.5                             | 2.3   | 1.7   |
| Mean absolute error (MAE) | 0.3                             | 0.3   | 0.4   |
| Standard deviation (SD)   | 0.2                             | 0.3   | 0.3   |
| Test set energy range     |                                 | 86.0  |   |
| Training set energy range |                                 | 114.9   |   |
| MAE% error                | 0.3                             | 0.4   | 0.4   |

All energies are given in  $\text{kJ mol}^{-1}$ . MAE% represents the MAE error with respect to the energy range of the test set

**Fig. 6** Breakdown of atomic energy errors per atom for methanol. All energies are in  $\text{kJ mol}^{-1}$ 

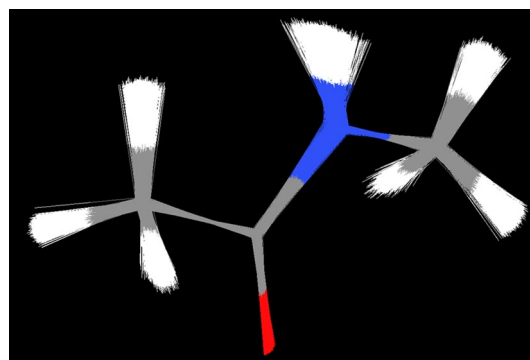
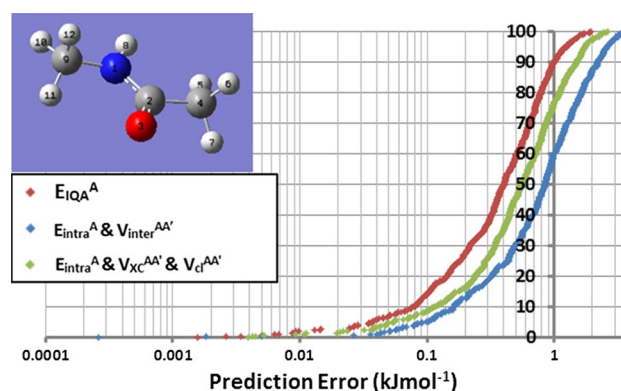
MAEs more comparable when assessing the difficulty for the kriging engine. The MAE percentage errors, along with the MAEs and energy ranges, are tabulated in Table S1 in the Electronic Supplementary Material. Approach A corresponds to S1 (a), approach B to S1 (b) and approach C to S1 (c). The analysis for the  $E_{\text{intra}}^A$  model [used in both approaches B and C and given in S1 (b)] is not repeated in S1 (c), as the same model is used for each approach. Figure S1 plots MAE versus atomic energy range in order to observe any correlation between them. Some weak correlation can be seen, but nothing strong enough to validate such a relationship.

## 4.2 NMA

The NMA models were trained using 3300 training set samples (Fig. 7), sampling an energy well with an energy range of  $\sim 84 \text{ kJ mol}^{-1}$ .

Figure 8 plots the S-curve for each NMA molecular model.

In Fig. 8, almost 95 % of the all test set samples have the  $E_{\text{IQA}}^{\text{Molec}}$  energy correctly predicted within  $2.5 \text{ kJ mol}^{-1}$  (across all models). This time there is a clearer separation of

**Fig. 7** Set of 4000 distorted NMA samples as generated from TYCHE through sampling of the normal modes at 450 K**Fig. 8** NMA S-curves showing the absolute errors for each of the three modelling approaches, each tested on the same 500 test set samples

the S-curve models, with again approach A ( $E_{\text{IQA}}^A$ ) being the best modelled of the 3 approaches. Interestingly and unexpectedly, approach C ( $E_{\text{intra}}^A$ ,  $V_{\text{XC}}^{A,A'}$  and  $V_{\text{cl}}^{A,A'}$ ) performs better than approach B ( $E_{\text{intra}}^A$  and  $V_{\text{inter}}^{A,A'}$ ) for the NMA system. Given this result, we must now consider whether the dual-cancellation (atomic and molecular) allowed in Eq. (13), prevents

**Table 2** Quantitative analysis of the NMA models

| Measure                   | $E_{\text{IQA}}^A$ (approach A) | $E_{\text{intra}}^A$ & $V_{\text{inter}}^{AA'}$ (approach B) | $E_{\text{intra}}^A$ & $V_{\text{XC}}^{AA'}$ & $V_{\text{cl}}^{AA'}$ (approach C) |
|---------------------------|---------------------------------|--|---|
| Maximum absolute error    | 1.9                             | 3.8  | 2.6   |
| Minimum absolute error    | 0.0                             | 0.0  | 0.0   |
| Absolute error range      | 1.9                             | 3.8  | 2.6   |
| Mean absolute error (MAE) | 0.5                             | 1.0  | 0.7   |
| Standard deviation (SD)   | 0.4                             | 0.8  | 0.5   |
| Test set energy range     |                                 | 81.1   |   |
| Training set energy range |                                 | 83.9   |   |
| MAE% error                | 0.6                             | 1.3  | 0.9   |

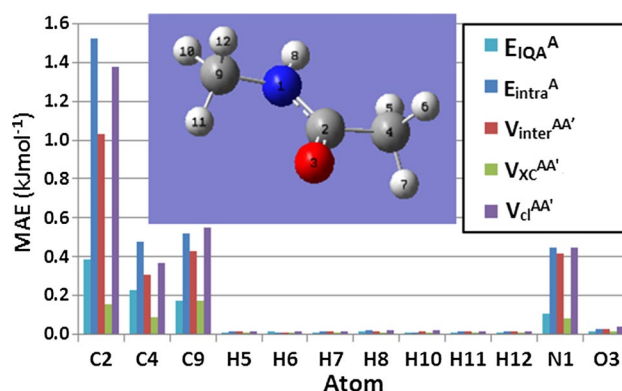
All energies are given in  $\text{kJ mol}^{-1}$ . MAE% represents the MAE error with respect to the energy range of the test set

a straightforward correlation between the number of atomic models composing a molecular model and the quality of the molecular model. Glycine will further aid our understanding of this, and the topic will be discussed further in the Sect. 5.

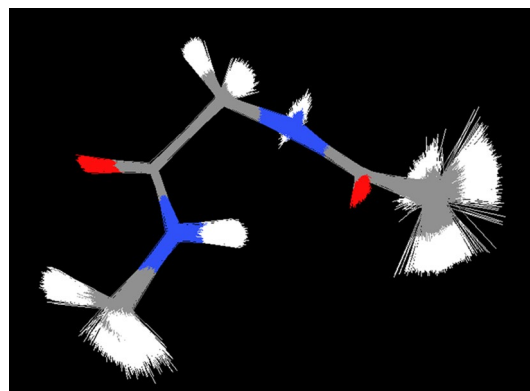
The MAE percentage errors for approaches A, B and C are 0.6, 1.3 and 0.9 %, respectively (see Table 2). Again, this is a pleasing result considering the molecular complexity has risen from  $(3 \times 6) - 6 = 12$  geometrical features for methanol, to  $(3 \times 12) - 6 = 30$  features for NMA. While the number of geometrical features increased by a factor 2.5, the errors for approaches A and C barely doubled. This favourable behaviour stimulates a further upscaling of features. With the *S*-curves being less entangled for NMA, the maximum absolute error falls in line with the shape and position of the respective *S*-curve.

Figure 9 is the counterpart of Fig. 6, this time for NMA. Here, we can further confirm that the atomic energy MAEs appear directly related to both the element and energy type being modelled. Initially, the atoms forming NMA can be immediately separated into their elements for the carbon, oxygen and hydrogen atoms, but the nitrogen atoms have very similar errors to the methyl carbons. In NMA, the MAEs also separate the atoms into atom types for carbon (carbonyl carbon and methyl-cap carbons are easily distinguishable). In fact, the oxygen is also modelled so well it is close to being indistinguishable from the hydrogens. Following this, the same trend in energy prediction accuracy is present in NMA as it was in methanol, that is, the sequence  $E_{\text{intra}}^A$  &  $V_{\text{cl}}^{AA'} < V_{\text{inter}}^{AA'} < E_{\text{IQA}}^A < V_{\text{XC}}^{AA'}$  (most accurate) remains valid.

The MAE percentage errors, MAEs and energy ranges are all reported for each atom in Table S2. Figure S3 once more confirms the weak correlation between MAE and energy range, this time for NMA. Figure S4 is analogous to Fig. 9, but plotting MAE% instead of MAE. In going from MAE to MAE%, the range of the energy is now incorporated. As we can see in Figure S4, the trend previously identified for the MAE ( $E_{\text{intra}}^A$  &  $V_{\text{cl}}^{AA'} < V_{\text{inter}}^{AA'} < E_{\text{IQA}}^A < V_{\text{XC}}^{AA'}$ ) is no longer true, and no clear trend is seen.



**Fig. 9** Breakdown of atomic energy errors per atom for NMA. Energies are in  $\text{kJ mol}^{-1}$



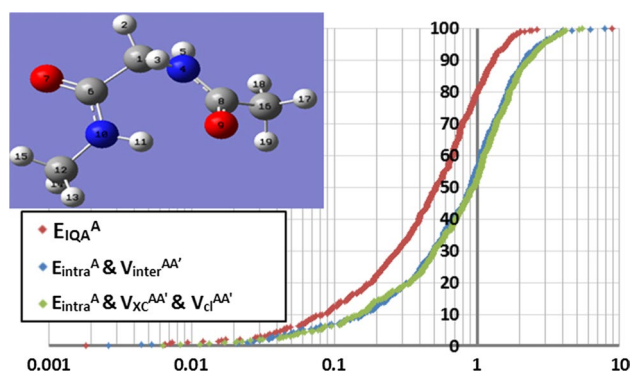
**Fig. 10** Set of 4000 distorted capped glycine samples generated from TYCHE through sampling of the normal modes at 450 K

### 4.3 Glycine (Gly)

The capped glycine models were trained using 3000 training set samples (Fig. 10), sampling an energy range of  $\sim 163 \text{ kJ mol}^{-1}$ .

Figure 11 shows the *S*-curve for glycine.





**Fig. 11** Capped glycine *S*-curves showing the absolute errors for each of the three modelling approaches, each tested on the same 500 test set samples

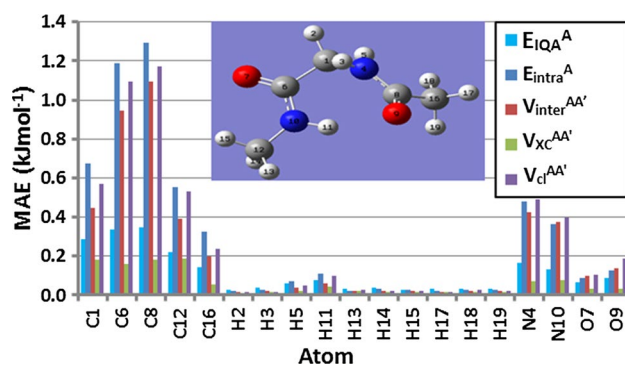
**Table 3** Quantitative analysis of the capped glycine models

| Measure                   | $E_{\text{IQA}}^A$ | $E_{\text{intra}}^A$ & $V_{\text{inter}}^{AA'}$ | $E_{\text{intra}}^A$ & $V_{\text{XC}}^{AA'}$ & $V_{\text{cl}}^{AA'}$ |
|---------------------------|--------------------|---|--|
| Maximum absolute error    | 8.9                | 7.9   | 5.5  |
| Minimum absolute error    | 0.0                | 0.0   | 0.0  |
| Absolute error range      | 8.9                | 7.9   | 5.5  |
| Mean absolute error (MAE) | 0.6                | 1.1   | 1.1  |
| Standard deviation (SD)   | 0.6                | 0.9   | 0.9  |
| Test set energy range     |                    | 162.5   |  |
| Training set energy range |                    | 178.0   |  |
| MAE% error                | 0.4                | 0.7   | 0.7  |

All energies are given in  $\text{kJ mol}^{-1}$ . MAE% represents the MAE error with respect to the energy range of the test set

The inset glycine conformation in Fig. 11 is the global minimum identified by TYCHE as forming  $\sim 75\%$  of the Boltzmann distribution when all 9 energy minima were used as seeds. The predominance of the global minimum determined our choice to sample only around this conformation. Almost 95 % of the all test set samples have the  $E_{\text{IQA}}^{\text{Molec}}$  energy correctly predicted within  $\sim 2.9 \text{ kJ mol}^{-1}$  (across all models). The *S*-curves show a resemblance to those in Fig. 5 but shifted to a higher prediction error. Here, there is no longer a clear separation of the *S*-curves for modelling approaches *B* and *C*. Approach *A* ( $E_{\text{IQA}}^A$ ) still has the lowest errors of the three approaches. Surprisingly approaches *B* ( $E_{\text{intra}}^A$  and  $V_{\text{inter}}^{AA'}$ ) and *C* ( $E_{\text{intra}}^A$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ ) is that they again result in the same MAEs of  $1.1 \text{ kJ mol}^{-1}$ . These data are presented in Table 3.

The MAE percentage errors for approaches *A*, *B* and *C* are 0.6, 0.9 and 0.9 %, respectively (see Table 3). Again, this is a very pleasing result considering the molecular complexity has once more risen from  $(3 \times 12) - 6 = 30$



**Fig. 12** Breakdown of atomic energy errors per atom for capped glycine. Energies are in  $\text{kJ mol}^{-1}$

geometrical features for NMA, to  $(3 \times 19) - 6 = 51$  features for glycine. In spite of a near doubling of the number of geometrical features, there is little change in the MAEs going from NMA to glycine.

Figure 12 reconfirms the trends identified in Fig. 6 and Fig. 9. Indeed, MAEs are related to the element being modelled and IQA energy type, with atom typing appearing even more evident. Within the carbons, three types are present but only two classes are distinguishable:  $\text{C}_{\text{C=O}} > \text{C}_{\alpha} \equiv \text{C}_{\text{methyl}}$ . Once more, the nitrogen atoms are fairly indistinguishable from the latter class of carbons. This time, the amino hydrogens (H5 and H11) also appear with slightly higher errors than seen for the aliphatic hydrogens. When looking at the trends in the energies themselves, the trends previously seen for methanol and NMA are once more observed in Gly, where  $E_{\text{intra}}^A$  &  $V_{\text{cl}}^{AA'} < V_{\text{inter}}^{AA'} < E_{\text{IQA}}^A < V_{\text{XC}}^{AA'}$  (most accurate). A direct comparison of the errors on the atoms present in both NMA and Gly will be given in the Sect. 5.

Table S3 and the corresponding plots of Figures S5 and S6 are analogous to Table S2 and Figures S3 and S4, this time for glycine. The same observation of a weak correlation between energy range and MAE for each IQA energy type is made in Figure S5. Figure S6 displays the MAE% values for capped glycine. It is clear that 13 out of 19 atoms show  $E_{\text{IQA}}^A$  standing out as the least accurate energy to model. For previous systems this majority trend was not seen. However, this conclusion can be rationalized by remembering that  $E_{\text{IQA}}^A$  is the total atomic energy, and hence influenced by every type of energy change within the atom. Hence, it would be reasonable for it to be the most sensitive when the energy is considered relative to the energy range.

## 5 Discussion

The discussion is divided into four subsections covering key topics that have either been postulated at the beginning

of the research or have arisen during the analysis of the results. The first two subsections each correspond to an objective.

### 5.1 Feasibility of modelling IQA energies

The first objective of this investigation is to assess the suitability and quality of modelling the five IQA energies that can be used in the formulation of FFLUX. In short, we have successfully kriged models and made good molecular energy predictions using three possible combinations of the IQA energies, making them all suitable for use in FFLUX. The resulting molecular models having excellent errors, of less than  $\pm 0.4$ , 1.3 and 0.7 kJ mol<sup>-1</sup> for methanol, NMA and Gly, respectively. Qualitatively speaking, Figs. 5, 8 and 11 display behaviour analogous to previous *S*-curves in previous literature [31, 39], leading to an overall successful prediction of both multipolar electrostatic and non-electrostatic energetics.

In order to quantitatively compare the quality of the results, we need to draw on a previous paper [39] where a component of  $E_{\text{intra}}^A$ , namely the kinetic energy  $T^A$ , was kriged for every atom in a similar set of systems (methanol, NMA, glycine and triglycine). We decided to compare our results with the kinetic energy results, only for methanol and NMA, because their training set sizes match best. In that work [39], MAEs for the atomic kinetic energy were obtained of 0.8 kJ mol<sup>-1</sup> (0.1 %) and 0.7 kJ mol<sup>-1</sup> (0.3 %) for a methanol-carbon and the carbonyl-carbon in NMA, respectively. In our work, we have presented differing training set sizes (3400 and 3300 for methanol and NMA, respectively), but it is still useful to compare the results. For  $E_{\text{intra}}^A$  we obtain MAEs of 0.7 kJ mol<sup>-1</sup> (0.3 %) and 1.5 kJ mol<sup>-1</sup> (0.4 %), respectively, and for  $V_{\text{inter}}^{AA'}$  we have MAEs of 0.5 kJ mol<sup>-1</sup> (0.3 %) and 1.0 kJ mol<sup>-1</sup> (0.2 %) for the equivalent atoms, respectively. Hence, despite our training sets being larger, the MAE errors remain slightly higher than those observed for the kinetic energy. This confirms an initial suspicion that the summative nature of both  $E_{\text{intra}}^A$  and  $V_{\text{inter}}^{AA'}$  results in a more complicated kriging problem than an example of the subcomponents ( $T^A$ ) forming these energies. However, our overall similar performance is still very promising given that we have the complete energy of an atom *A* (and thus a molecule when summing over *A*) being modelled with comparable errors, albeit using larger training sets. Another advantage of this investigation is the ability to krig only one, two or three energies, yet still capturing the energetic behaviour of the whole molecule. Hence, this design saves substantial computational time by not needing to krig every individual IQA atomic energy ( $T^A$ ,  $V_{\text{en}}^{AA}$ ,  $V_{\text{ee}}^{AA}$ ,  $V_{\text{en}}^{AA'}$ ,  $V_{\text{ne}}^{AA'}$ ,  $V_{\text{nn}}^{AA'}$  and  $V_{\text{ee}}^{AA'}$ ) (should  $V_{\text{ee}}$  still remain unpartitioned). It is also noted that across all atoms in all three systems investigated here, the MAE

error never exceeded 1.5 kJ mol<sup>-1</sup> (with the majority under 1 kJ mol<sup>-1</sup>), or a MAE percentage error over 1.4 %, for any energy.

Another measure of quality to loosely compare our results with are the previously kriged electrostatic multipole moments, which describe the classical electrostatic interaction energy for 1,4 and higher order interactions [33]. Here, the notation ‘1,4’ describes the interaction between atoms separated by 3 covalent bonds. A 1,5 interaction has 4 separating covalent bonds, and so on. For 1,4 and higher interactions (i.e. 1,  $n > 4$ ) in a capped histidine system (29 atoms), the MAE for the intramolecular electrostatic energy calculated through kriged multipole moments, was 2.5 kJ mol<sup>-1</sup>. In our investigation  $V_{\text{cl}}^{AA'}$  and  $V_{\text{inter}}^{AA'}$  never exceed an MAE of 1.4 kJ mol<sup>-1</sup>. Is this MAE respectable compared to the multipolar electrostatic energy error? Yes, because  $V_{\text{cl}}^{AA'}$  (and  $V_{\text{inter}}^{AA'}$  too) accounts for all electrostatic interactions and the multipolar electrostatic analysis only for 1,4 and higher interactions (due to convergence limitations). Admittedly, a training set of only 600 training set geometries [33] was used for the latter analysis.

Finally, we point out that we are currently investigating a potential reduction in the number of training set samples needed to obtain suitably accurate atomic and molecular models. This research is focussed on the selective building of training sets, and a variety of approaches are currently being investigated to achieve this.

### 5.2 Cancellation of errors

A second objective of the current investigation is to observe to what extent any cancellation of errors takes place within the summative combination of kriging models described in Eq. 9. As previously described, this approach offers the potential to benefit from the fortuitous cancellation of errors, but also equally the unfortunate accumulation of errors as a result of the machine learning. In particular, an atomic energy component may be, on average, predicted to be more stable than the true energy (i.e. *overestimated*). Another atomic energy component may be, on average, predicted to be less stable than the true energy (i.e. *underestimated*). As a result, the summative combination of both an over- and underestimated result allows for some cancellation, resulting in an overall energy prediction being more accurate, when only these two energies are considered. Accumulating these cancellations further across many energy models and across many atoms increases this effect dramatically.

Despite the cancellation appearing to rely on chance, since there is no control for the over- or underestimation of the energy models, the prediction of the kriging models will always be consistent, i.e. the same geometrical features are used to map all atomic energies within a single atom,

and each atom's geometrical features are related to every other atom in the molecule. Hence, if one energy is overestimated and the another is underestimated with those identical or related features, then this same interplay between the kriging models will always be present. Naturally, the opposite can also occur where, for example, two overestimating models result in a summed higher total error instead of cancelling. However, from the results in Figs. 6, 9 and 12, it is evident that the summative energies ( $E_{\text{IQA}}^A$  and  $E_{\text{inter}}^{AA'}$ ) generally have lower errors than by simply summing the absolute errors of the components that make up these energies ( $E_{\text{intra}}^A + E_{\text{inter}}^{AA'}$  and  $V_{\text{XC}}^{AA'} + V_{\text{cl}}^{AA'}$ , respectively). Evidence of the molecular cancellation (described in Sect. 3.2) can also be seen, given that atomic MAE often are between 0.5 and 1.5 kJ mol<sup>-1</sup>, but the resulting molecular MAE is always  $\leq 1$  kJ mol<sup>-1</sup> (for NMA).

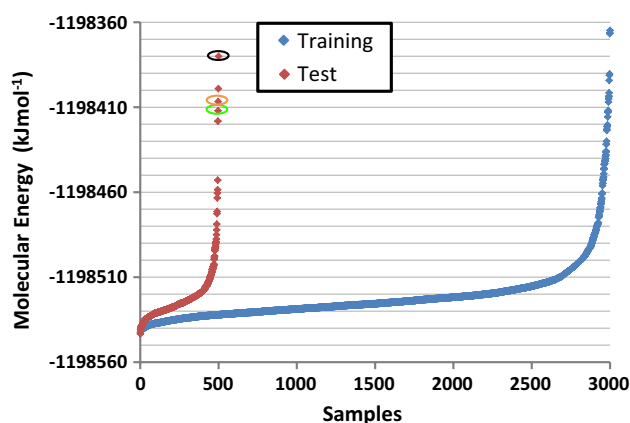
### 5.3 S-curve analysis

The MAEs (and MAE percentage errors) of the total  $E_{\text{IQA}}^{\text{Molec}}$  molecular energy, as given in Tables 1, 2 and 3 for each respective molecule, are arguably the most important values obtained in this investigation. Therefore, these values are representative of an overall quality check for this investigation. Through averaging the MAE of the three approaches (A, B and C) for each molecule, we obtain a single MAE error for each molecule: 0.3 kJ mol<sup>-1</sup> (methanol), 0.7 kJ mol<sup>-1</sup> (NMA) and 0.9 kJ mol<sup>-1</sup> (Gly). These  $<1$  kJ mol<sup>-1</sup> results become more impressive when the total energy for each system is compared: methanol has a molecular energy of  $\sim 302,000$  kJ mol<sup>-1</sup>, NMA  $\sim 648,700$  kJ mol<sup>-1</sup> and Gly  $\sim 1,191,500$  kJ mol<sup>-1</sup>.

Of the three approaches, approach A consistently is the most accurate for each molecule. This is a result of the minimal approach incorporating only a single model per atom in the molecule. However, the performance of approaches B and C were less distinguishable or predictable. Either approach was capable of being slightly more accurate than the other. However, the molecular error when calculated using either approach B or C was always within 0.3 kJ mol<sup>-1</sup> of one another. In conclusion, we consider all three routes as suitable candidates for modelling the molecular energy, each incorporating a different level of chemical insight.

One further point to note in the analysis of our S-curves, are the “rogue points” present near the 100 % ceiling of the plots. Few rogue points occur for approach B in methanol, but more noticeably for approaches B and, in particular, approach A for Gly. These points are considered rogue due to the large gap that appears separating these points from the almost continuous S-like shape of the plot. Figure 13 (on capped glycine) sheds lights on how rogue points arise.

In Fig. 13, the glycine geometries that passed GAIA's scrubbing procedure are plotted according to their



**Fig. 13** Set of 3000 training samples and 500 test set samples, with randomly assigned sample numbers (*x*-axis), plotted against their molecular energies (*y*-axis) for Gly. The Rogue Test Point 1 (RTP1) is encircled green, RTP2 orange, and the Test Point TP1 black

molecular energies and separated according to which set they belong to, that is, the training set in blue or the test set in red. The information shown in Fig. 13 is essentially a one-dimensional distribution of molecular energies (*y*-axis) but spread out in two dimensions by introducing an *x*-axis that merely counts the 3000 training set samples and the 500 test set samples. The test set remains identical for all three modelling approaches.

A large vertical gap between the blue points in Fig. 13 indicates a lack of training points in that energy region. On the other hand, continuous lines indicate a high density of points, covering well the corresponding energy region. The point in the test set at  $-1,198,412.1$  kJ mol<sup>-1</sup> (encircled green in Fig. 13) is the geometry corresponding to the maximum predicted error (8.9 kJ mol<sup>-1</sup>) seen in the S-curve of approach A (utmost right point in the red curve in Fig. 11), denoted RTP1 (Rogue Test Point 1). The point in the test set at  $-1,198,406.7$  kJ mol<sup>-1</sup> (encircled orange in Fig. 13) is the geometry corresponding to the maximum error (7.9 kJ mol<sup>-1</sup>) seen in the S-curve of approach B (utmost right point in blue curve in Fig. 11), denoted RTP2. Both these molecular energies appear to be reasonably well sampled in the training set, with nearby (blue) points of  $-1,198,412.5$  and  $-1,198,406.9$  kJ mol<sup>-1</sup>. Unusually, there was no problem for approaches B and C in predicting the molecular energy for RTP1, with errors of 3.4 and 1.3 kJ mol<sup>-1</sup>, respectively. Similarly, for RTP2, errors of 1.1 and 3.1 kJ mol<sup>-1</sup> were obtained for approaches A and C. This suggests that it is generally not a lack of training geometries that are causing the large rogue errors. Instead it could be any one (or combination) of the following three effects: (1) the potential energy surface around these points is undulant (in general for the molecule, or for a particular IQA energy) and the training geometries included are

not enough to fully capture this landscape adequately, (2) the test points are outside of the domain of applicability, defined as the region of conformational space that can be interpolated by the training points of the kriging model. In other words, points lying outside of the domain of applicability correspond to points that lie outside of the training set, and so the kriging model is required to perform an extrapolation to make a prediction [74], (3) when summing across models, the balance of accumulating errors results in reduced cancellation of errors. All these reasons would also be supported by working with the higher-energy geometries where there are fewer samples in the training set, compared to those closer to the energy of the seed minimum.

Another measure that is useful when analysing the cause of a poor prediction within FEREBUS is the mean signed error (MSE) (or mean signed deviation, MSD). In statistics, the MSE is a measure of how close a predicted value matches the true quantity. Having a high MSE for a particular prediction indicates that the model is not well trained for in that region, and is a hallmark of working outside the domain of applicability. Taking glycine's approach A as an example, the C12 atom stands out as an atom with a particularly poor  $E_{\text{IQA}}^A$  prediction for RTP1 (an error of  $5.2 \text{ kJ mol}^{-1}$ ). C12 also has an MSE approximately five times the average across all of the test geometries. Some of the other atoms in glycine also indicate a slightly increased difficulty in predicting  $E_{\text{IQA}}^A$  for this test geometry, but not to the same degree as for C12. Hence, it can be concluded that C12 is the source of the error for RTP1, due to the model operating outside of its domain of applicability. Evidently, as approaches B and C perform well in predicting this molecular energy, this MSE explanation either is irrelevant when using  $E_{\text{intra}}^A$ ,  $V_{\text{inter}}^{AA'}$ ,  $V_{\text{cl}}^{AA'}$ , and/or  $V_{\text{XC}}^{AA'}$ , or the effect is dampened by the cancellation of errors. This type of analysis can be applied to any rogue point on an *S*-curve.

In contrast, the point at  $-1,198,380.1 \text{ kJ mol}^{-1}$  (encircled black in Fig. 13) in the test set (TP1) (which is not a rogue point), appears to be the least well sampled in the training set, but the predicted errors for this sample are  $0.3 \text{ kJ mol}^{-1}$  (approach A),  $2.4 \text{ kJ mol}^{-1}$  (approach B) and  $2.0 \text{ kJ mol}^{-1}$  (approach C), thus, not near the maximal points on each of the *S*-curves. The unexpectedly good prediction of TP1 is credited to kriging's impressive interpolation between two largely spaced training points.

## 5.4 Evidence for atom typing

Figures 6, 9 and 12 illustrated the 'difficulty' of modelling each atomic energy, according to the MAE. From this analysis, we learned that atoms belonging to a particular functional group had a MAE that distinguished some from others, independently of the IQA energies being used for this observation. Across our three systems, the carbonyl carbons

were the most difficult atoms to model, with a maximum MAE value of  $\sim 1.2\text{--}1.3 \text{ kJ mol}^{-1}$ , in both NMA and Gly. The carbonyl carbon was followed by similar maximum MAEs values for the  $\alpha$ -carbon, the amino nitrogens and methyl carbons of  $\sim 0.5 \text{ kJ mol}^{-1}$ , in methanol, NMA and Gly. The oxygens were easily distinguishable with a much lower maximum MAE of around  $0.1 \text{ kJ mol}^{-1}$ , followed by the consistently very accurately modelled hydrogens with maximal MAEs of  $<0.1 \text{ kJ mol}^{-1}$ . It is interesting to note that the N1 atom in NMA, resulted in similar errors to that of the corresponding atoms N4 and N10 in Gly. The same is true of the NMA atoms carbonylic C2 (with C6 and C8 in Gly) and O3 (with O7 and O9 in Gly), within  $\pm 0.2 \text{ kJ mol}^{-1}$ . Hence, with atoms having comparable MAEs across multiple molecules, there is some basic evidence of atom typing. However, it is not a rule that can be used in distinguishing all present functional groups, as evidenced by the difficulty in distinguishing  $\alpha$ -carbon, methyl carbon and the amino nitrogen groups, using only the MAEs. It would be interesting to see which further trends are observed when a broader range of functional groups are studied.

The discussion above is not the first time atom typing has been considered within the study of an energy partitioning. A recent article by Patrikeev et al. [75], investigated the performance of several density functionals in their evaluation of both Kohn–Sham and correlation kinetic energies of topological atoms, and also commented on discriminating atom types through such atomic descriptors. Initial findings for the Kohn–Sham energies indicated a strong link between some of the tested functionals and the atomic number (or element) of an atom. A further finding in the assessment of correlation kinetic energies allowed aromatic and aliphatic hydrogens to be separated. It should also be reiterated that IQA within DFT is a little tricky since the Kohn–Sham approach does not lead to exact correlated reduced density matrices [76].

## 6 A note on dispersion and transferability

The only type of energy contribution that is lacking in the current kriging treatment of all IQA energy contributions is that associated with dispersion. Admittedly, the current treatment includes electron correlation, but because we used B3LYP this electron correlation does not cover dispersion effects. However, soon-to-be-published work of our group successfully kriges the IQA intra-atomic,  $E_{\text{intra}}^A$ , and interatomic,  $V_{\text{inter}}^{AA'}$ , energies at the M06-2X/aug-cc-pVDZ level of theory. This functional describes (or mimics) some mid-range dispersion effects, but the ultimate goal of FFLUX is to invoke a post-Hartree–Fock method (non-DFT) to cover dispersion properly.



A second note concerns the transferability of the obtained model with respect to an exchange–correlation functional other than B3LYP. The only other exchange–correlation functionals implemented in the program AIM-ALL are LSDA and M06-2X. The current investigation has not been mirrored using any other functional from which we could directly compare transferability results. However, there is potential for transferability to be considered in the aforementioned work to be published, where M06-2X was used. Like in the current work, IQA atomic energy predictions made for multiple different systems could be compared. Some understanding of the transferability of another exchange–correlation functional can come from earlier work [77] from our laboratory in which the electrostatic energy, obtained through atomic multipole moments, was kriged at three levels of theory, namely HF, B3LYP and M06-2X. From that work, one would expect that M06-2X will perform similarly to B3LYP.

## 7 Conclusion

The development of the novel force field FFLUX now moves beyond its machine learning treatment of multipolar electrostatics. We demonstrate that short-range non-multipolar electrostatics can now also be kriged successfully. Moreover, (non-multipolar) exchange energies as well as intra-atomic energies (beyond just the kinetic energy) are now also kriged with promising energy errors. As a result, chemical bonding and stereo-electronic effects are now, by way of principle, incorporated in FFLUX. This achievement is realized within the context of the methodology of interacting quantum atoms (IQA).

Three approaches (A, B, and C), incorporating five IQA atomic energies ( $E_{\text{IQA}}^A$ ,  $E_{\text{intra}}^A$ ,  $V_{\text{inter}}^{AA'}$ ,  $V_{\text{XC}}^{AA'}$  and  $V_{\text{cl}}^{AA'}$ ), were successfully used to develop molecular models offering control in balancing accuracy and chemical insight. The most accurate and least expensive molecular model (approach A) was built using the total atomic energy  $E_{\text{IQA}}^A$ , with MAEs of  $\pm 0.3$ , 0.4 and 0.6 kJ mol<sup>−1</sup> for methanol, NMA and capped glycine, respectively. Interestingly, the more insightful formalisms involving the intra- and inter-atomic components (approach B), and the interatomic exchange and electrostatic contributions (approach C), resulted in similar MAEs. These errors are on a par with previous literature and are a result of the combination of models benefitting from cancellation of errors. The latter occur both within an atom's total energy modelling (*atomic* cancellation), and also when summing *across* total atomic models (*molecular* cancellation) to obtain the molecular model.

In summary, the novel strategy and results were a successful proof-of-concept approach, developed to be

integrated into FFLUX. Future work will build upon the method presented here employing the models in the application of geometry optimization, initially without, but later with, the incorporation of multipolar electrostatics. Future research is also focussing on creating intelligent training sets, designed to reduce the number of samples used in the building of kriging models.

**Acknowledgments** We acknowledge the EPSRC for funding through the award of an Established Career Fellowship to P.L.A.P. (Grant EP/K005472), and the BBSRC for studentships to S.C. and P.M.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? *Biophys J* 100:L47–L49
2. Rauscher S, Gapsys V, Gajda MJ, Zweckstetter M, de Groot BL, Grubmüller H (2015) Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J Chem Theor Comput* 11:5513–5524
3. Ponder JW, Wu C, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RAJ, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T (2010) Current status of the AMOEBA polarizable force field. *J Phys Chem B* 114:2549–2564
4. Vinter JG (1994) Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. *J Comput Aided Mol Des* 8:653–668
5. Gresh N, Cisneros GA, Darden TA, Piquemal J-P (2007) Anisotropic, polarizable molecular mechanics studies of inter- and intramolecular interactions and ligand-macromolecule complexes. A bottom-up strategy. *J Chem Theory Comput* 3:1960–1986
6. Verstraelen T, Vandenbrande S, Ayers PW (2014) Direct computation of parameters for accurate polarizable force fields. *J Chem Phys* 141:194114
7. Cardamone S, Hughes TJ, Popelier PLA (2014) Multipolar electrostatics. *Phys Chem Chem Phys* 16:10367–10387
8. Kramer C, Spinn A, Liedl KR (2014) Charge anisotropy: where atomic multipoles matter most. *J Chem Theory Comput* 10:4488–4496
9. Kosov DS, Popelier PLA (2000) Convergence of the multipole expansion for electrostatic potentials of finite topological atoms. *J Chem Phys* 113:3969–3974
10. Popelier PLA, Joubert L, Kosov DS (2001) Convergence of the electrostatic interaction based on topological atoms. *J Phys Chem A* 105:8254–8261
11. Popelier PLA, Kosov DS (2001) Atom-atom partitioning of intramolecular and intermolecular Coulomb energy. *J Chem Phys* 114:6539–6547
12. Popelier PLA, Rafat M (2003) The electrostatic potential generated by topological atoms: a continuous multipole method leading to larger convergence regions. *Chem Phys Lett* 376:148–153

13. Rafat M, Popelier PLA (2005) The electrostatic potential generated by topological atoms. Part II: inverse multipole moments. *J Chem Phys* 123(204103–204101):204107
14. Rafat M, Popelier PLA (2006) A convergent multipole expansion for 1,3 and 1,4 Coulomb interactions. *J Chem Phys* 124(144102):1–7
15. Rafat M, Popelier PLA (2007) Topological atom-atom partitioning of molecular exchange energy and its multipolar convergence. In: Matta CF, Boyd RJ (eds) *Quantum theory of atoms in molecules*, vol 5. Wiley-VCH, Weinheim, pp 121–140
16. Rafat M, Popelier PLA (2007) Long range behaviour of high-rank topological multipole moments. *J Comput Chem* 28:832–838
17. Joubert L, Popelier PLA (2002) The prediction of energies and geometries of hydrogen bonded DNA base-pairs via a topological electrostatic potential. *Phys Chem Chem Phys* 4:4353–4359
18. Bader RFW (1990) *Atoms in molecules. A quantum theory*. Oxford University Press, Oxford
19. Popelier PLA (2000) *Atoms in molecules. An introduction*. Pearson Education, London
20. Popelier PLA (2014) The quantum theory of atoms in molecules, Chapter 8. In: Frenking G, Shaik S (eds) *The nature of the chemical bond revisited*. Wiley-VCH, Weinheim, pp 271–308
21. Matta CF, Boyd RJ (2007) *The quantum theory of atoms in molecules. From solid state to DNA and drug design*. Wiley-VCH, Weinheim
22. Popelier PLA (2015) QCTFF: on the construction of a novel protein force field. *Int J Quantum Chem* 115:1005–1011
23. Bader RFW, Popelier PLA (1993) Atomic theorems. *Int J Quantum Chem* 45:189–207
24. Yuan Y, Mills MJL, Popelier PLA (2014) Multipolar electrostatics for proteins: atom–atom electrostatic energies in crambin. *J Comput Chem* 35:343–359
25. Handley CM, Popelier PLA (2010) Potential energy surfaces fitted by artificial neural networks. *J Phys Chem A* 114:3371–3383
26. Handley CM, Popelier PLA (2009) A dynamically polarizable water potential based on multipole moments trained by machine learning. *J Chem Theory Comput* 5:1474–1489
27. Handley CM, Hawe GI, Kell DB, Popelier PLA (2009) Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning. *Phys Chem Chem Phys* 11:6365–6376
28. Cressie N (1993) *Statistics for spatial data*. Wiley, New York
29. Rasmussen CE, Williams CKI (2006) *Gaussian processes for machine learning*. The MIT Press, Cambridge
30. Rupp M, Ramakrishnan R, von Lilienfeld OA (2015) Machine learning for quantum mechanical properties of atoms in molecules. *J Phys Chem Lett* 6:3309–3313
31. Fletcher TL, Davie SJ, Popelier PLA (2014) Prediction of intramolecular polarization of aromatic amino acids using kriging machine learning. *J Chem Theory Comput* 10:3708–3719
32. Fletcher TL (2014) Polarizable multipolar electrostatics driven by kriging machine learning or a peptide force field: assessment, improvement and up-scaling. Ph.D. Thesis, School of Chemistry, University of Manchester
33. Kandathil SM, Fletcher TL, Yuan Y, Knowles J, Popelier PLA (2013) Accuracy and tractability of a kriging model of intramolecular polarizable multipolar electrostatics and its application to histidine. *J Comput Chem* 34:1850–1861
34. Yuan Y, Mills MJL, Popelier PLA (2014) Multipolar electrostatics based on the kriging machine learning method: an application to serine. *J Mol Model* 20:2172–2186
35. Mills MJL, Popelier PLA (2012) Polarisable multipolar electrostatics from the machine learning method kriging: an application to alanine. *Theor Chem Acc* 131:1137–1153
36. Yuan Y (2012) A polarisable multipolar force field for peptides based on kriging: towards application in protein crystallography and enzymatic reactions. Ph.D. thesis, School of Chemistry, University of Manchester
37. Mills MJL, Popelier PLA (2011) Intramolecular polarisable multipolar electrostatics from the machine learning method kriging. *Comput Theor Chem* 975:42–51
38. Darley MG, Popelier PLA (2008) Role of short-range electrostatics in torsional potentials. *J Phys Chem A* 112:12954–12965
39. Fletcher TL, Kandathil SM, Popelier PLA (2014) The prediction of atomic kinetic energies from coordinates of surrounding atoms using kriging machine learning. *Theor Chem Acc* 133(1499):1410–1491
40. Blanco MA, Martin Pendas A, Francisco E (2005) Interacting Quantum Atoms: a correlated energy decomposition scheme based on the quantum theory of atoms in molecules. *J Chem Theor Comput* 1:1096–1109
41. Maxwell P, Martin Pendas A, Popelier PLA (2016) Extension of the interacting quantum atoms (IQA) approach to B3LYP level density functional theory. *Phys Chem Chem Phys* [**Epub ahead of print**]
42. Rafat M, Devereux M, Popelier PLA (2005) Rendering of quantum topological atoms and bonds. *J Mol Graph Model* 24:111–120
43. Rafat M, Popelier PLA (2007) Visualisation and integration of quantum topological atoms by spatial discretisation into finite elements. *J Comput Chem* 28:2602–2617
44. Silvi B, Savin A (1994) Classification of chemical bonds based on topological analysis of electron localization functions. *Nature (London)* 371:683–686
45. Popelier PLA (2005) Quantum chemical topology: on bonds and potentials. In: Wales DJ (ed) *Structure and bonding. Intermolecular forces and clusters*, vol 115. Springer, Heidelberg, pp 1–56
46. Popelier PLA (2016) Molecular simulation by knowledgeable quantum atoms. *Phys Scr* 91:033007
47. Popelier PLA, Aicken FM (2003) Atomic properties of selected biomolecules: quantum topological atom types of carbon occurring in natural amino acids and derived molecules. *J Am Chem Soc* 125:1284–1292
48. Popelier PLA (2016) Quantum chemical topology. In: Mingos DMP (ed) *The chemical bond—100 years old and getting stronger*. Springer, Cham, pp 71–117
49. Popelier PLA (2012) New insights in atom–atom interactions for future drug design. *Curr Top Med Chem* 12:1924–1934
50. Schneider N, Lange G, Hindle S, Klein R, Rarey M (2013) A consistent description of HYdrogen bond and DEhydration energies in protein–ligand complexes: methods behind the HYDE scoring function. *J Comput Aided Mol Des* 27:15–29
51. Bader RFW, Beddall PM (1972) Virial field relationship for molecular charge distributions and the spatial partitioning of molecular properties. *J Chem Phys* 56:3320–3329
52. Todd A Keith, AIMAll (Version 15.09.12) (2015) TK Gristmill Software, Overland Park KS, USA. <http://aim.tkgristmill.com>
53. Chávez-Calvillo R, García-Revilla M, Francisco E, Martín-Pendás A, Rocha-Rinza T (2015) Dynamical correlation within the interacting quantum atoms method through coupled cluster theory. *Comput Theor Chem* 1053:90–95
54. Eskandari K, Van Alsenoy C (2014) Hydrogen–hydrogen interaction in planar biphenyl: a theoretical study based on the interacting quantum atoms and Hirshfeld atomic energy partitioning methods. *J Comput Chem* 35:1883–1889
55. Dillen J (2013) Congested molecules. Where is the steric repulsion? An analysis of the electron density by the method of interacting quantum atoms. *Int J Quantum Chem* 113:2143–2153

56. Martin Pendas A, Blanco MA, Francisco E (2006) The nature of the hydrogen bond: a synthesis from the interacting quantum atoms picture. *J Chem Phys* 125:184112
57. Martin Pendas A, Francisco E, Blanco MA (2006) Binding energies of first row diatomics in the light of the interacting quantum atoms approach. *J Phys Chem A* 110:12864–12869
58. Inostroza-Rivera R, Yahia-Ouahmed M, Tognetti V, Joubert L, Herrera B, Toro-Labbe A (2015) Atomic decomposition of conceptual DFT descriptors: application to proton transfer reactions. *Phys Chem Chem Phys* 17:17797–17808
59. Di Pasquale N, Davie SJ, Popelier PLA (2016) Optimization algorithms in optimal predictions of atomistic properties by kriging. *J Chem Theor Comput* 12:1499–1513
60. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: *Proceedings of the IEEE international conference on neural networks*, vol 4, pp 1942–1948
61. Popelier PLA (2012) Quantum chemical topology: knowledgeable atoms in peptides. In: *AIP conference proceedings*, vol 1456, pp 261–268
62. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. *Proteins Struct Funct Genet* 17:412–425
63. Balsera MA, Wriggers W, Oono Y, Schulten K (1996) Principal component analysis and long time protein dynamics. *J Phys Chem* 100:2567–2572
64. Brooks B, Karplus M (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci USA* 80:6571–6575
65. Wilson D, Decius J, Paul C (1955) *Cross, molecular vibrations*. McGraw-Hill, New York
66. Ochterski JW (1999) Vibrational analysis in Gaussian. *Vibrational Analysis in Gaussian*. [http://www.gaussian.com/g\\_white-pap/vib.htm](http://www.gaussian.com/g_white-pap/vib.htm)
67. Hughes TJ, Cardamone S, Popelier PLA (2015) Realistic sampling of amino acid geometries for a multipolar polarizable force field. *J Comput Chem* 36:1844–1857
68. Yuan Y, Mills MJL, Popelier PLA, Jensen F (2014) Comprehensive analysis of energy minima of the 20 natural amino acids. *J Phys Chem A* 118:7876–7891
69. Jensen F (2002) Polarization consistent basis sets. III. The importance of diffuse functions. *J Chem Phys* 117:9234–9240
70. Martin Pendas A, Francisco E, Blanco MA, Gatti C (2007) Bond paths as privileged exchange channels. *Chem Eur J* 13:9362–9371
71. Francisco E, Martin Pendas A, Blanco MA (2006) A molecular energy decomposition scheme for atoms in molecules. *J Chem Theor Comput* 2:90–102
72. Martin Pendas A, Blanco MA, Francisco E (2007) Chemical fragments in real space: definitions, properties, and energetic decompositions. *J Comput Chem* 28:161–184
73. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JJA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam NJ, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) *Gaussian 09*. Gaussian Inc., Wallingford
74. Weaver S, Gleeson MP (2008) The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model* 26:1315–1326
75. Patrikeev L, Joubert L, Tognetti V (2016) Atomic decomposition of Kohn–Sham molecular energies: the kinetic energy component. *Mol Phys* 114:1285–1296
76. Tognetti V, Joubert L (2014) Density functional theory and Bader's atoms-in molecules theory: towards a vivid dialogue. *Phys Chem Chem Phys* 16:14539–14550
77. Hughes TJ, Kandathil SM, Popelier PLA (2015) Accurate prediction of polarised high order electrostatic interactions for hydrogen bonded complexes using the machine learning method kriging. *Spectrochim Acta A* 136:32–41